


- ORIGINAL ARTICLE -

# Machine Learning for Site Adaptation of Satellite-Derived Solar Irradiance in Northwestern Argentina.

Aprendizaje automático para la adaptación al sitio de irradiancia solar derivada de satélites en el Noroeste Argentino

Rubén Ledesma<sup>1,2</sup>  and Germán Salazar<sup>1,3</sup> 

<sup>1</sup>INECO - CONICET, Argentina

rdledesma@exa.unsa.edu.ar german.salazar@conicet.gov.ar

<sup>2</sup>Department of Informatic, Universidad Nacional de Salta, Argentina

<sup>3</sup>Department of Physics, Universidad Nacional de Salta, Argentina

## Abstract

Accurate estimation of global horizontal irradiance (GHI) is essential for solar energy resource assessment, particularly in regions with limited ground-based measurements. This study evaluates the performance of three machine learning models—Simple Linear Regression (SLR), Extreme Gradient Boosting (XGB), and Multilayer Perceptron (MLP)—for the site adaptation of satellite-derived GHI data in five locations in Northwestern Argentina. Two satellite products, CAMS and LSA-SAF, were used as input data. The models were assessed using standard error metrics (MBE, MAE, RMSE), and their residual patterns were analyzed. Results show that LSA-SAF data led to lower errors compared to CAMS, especially in high-altitude sites. While complex models like MLP and XGB marginally improved accuracy in some cases, SLR offered comparable results with higher robustness. The analysis also identified systematic biases and discretization effects in tree-based models. These findings suggest that, under current data conditions, simpler models may offer reliable performance. Enhancing input data quality and incorporating additional meteorological features may yield greater improvements than increasing model complexity.

**Keywords:** Solar irradiance, Site adaptation, Machine learning.

## Resumen

La estimación precisa de la irradiancia global horizontal (GHI) es fundamental para evaluar el recurso solar, especialmente en regiones con escasas mediciones terrestres. Este estudio evalúa el desempeño de tres modelos de aprendizaje automático—Regresión Lineal Simple (SLR), Extreme Gradient Boosting (XGB) y Perceptrón Multicapa (MLP)—para la adaptación al sitio de datos de GHI obtenidos por satélite en cinco ubicaciones del noroeste argentino. Se utilizaron dos productos satelitales, CAMS y LSA-SAF, como

datos de entrada. Los modelos se evaluaron mediante métricas estándar (MBE, MAE, RMSE) y análisis de residuos. Los resultados indican que los datos de LSA-SAF generaron errores menores, especialmente en sitios de gran altitud. Aunque modelos complejos como MLP y XGB mejoraron levemente la precisión en algunos casos, SLR logró resultados comparables con mayor robustez. El análisis también evidenció sesgos sistemáticos y efectos de discretización en modelos basados en árboles. Estos hallazgos sugieren que, dadas las condiciones actuales de los datos, modelos simples pueden ofrecer un desempeño confiable. Mejoras significativas podrían lograrse mediante la incorporación de variables meteorológicas adicionales y datos de mayor calidad.

**Palabras claves:** Irradiancia solar, Adaptación al Sitio, Aprendizaje automático

## 1 Introduction

Accurate knowledge of the solar resource distribution is crucial in the financial feasibility analysis of medium- or large-scale solar energy projects [1]. The Argentine Northwest (NOA) is among the regions of the world that receive above-average solar irradiation levels (Solargis.com). However, the spatial distribution of this solar resource across the region is not uniform [2]. In the NOA region, there are no regional radiometric networks that perform continuous measurements following established maintenance and recalibration protocols [3]. As a result, current solar radiation maps are not derived from ground-based measurement networks but rather from satellite-based estimations [4, 5]. Studies have shown that satellite-based models and reanalysis estimates exhibit differences when compared to ground-based measured values.

It is possible to adapt modeled solar irradiance values using ground-based measurements, provided that a time series of concurrent satellite-modeled and ground-

measured data is available at a specific site [6, 7]. The objective is to identify a function that adjusts — that is, improves — satellite-based estimates to more closely match actual ground measurements. This method is commonly referred to as Site Adaptation (SA).

Various Machine Learning (ML) models have been investigated to determine the optimal function to adapt these estimates [8, 9, 10]. The first results on SA for the NOA región were reported in [11, 3] and were consistent with the results obtained in related works.

The referenced studies employ models with varying levels of complexity and computational cost, ranging from simple approaches such as Simple Linear Regression to more sophisticated ones such as XGBoost and Multilayer Perceptron Artificial Neural Networks. These works reported commonly used performance metrics in the field, including Mean Bias Error (MBE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), which enabled them to identify the machine learning model that yields the best results.

However, limited attention has been given in previous studies to the underlying reasons why certain ML models outperform others in the solar assessment (SA) process. The objective of this study is to evaluate the SA process in five sites located in the NOA, encompassing the majority of available GHI estimation models. Furthermore, this work aims to provide a comprehensive analysis that documents the practical relevance and performance characteristics of each ML model employed.

## 2 Materials and Methods

### 2.1 Estimation models

#### 2.1.1 CAMS Heliosat-4

Heliosat-4 [12] is a physically based model that offers a fast and accurate alternative to the libRadtran radiative transfer model [13]. It integrates multiple satellite datasets along with reanalysis information from the Copernicus Atmosphere Monitoring Service (CAMS). Cloud characteristics are extracted from Meteosat Second Generation (MSG) imagery every 15 minutes using a modified version of the APOLLO algorithm (AVHRR3 Processing Scheme Over Clouds, Land, and Ocean) [14]. The model operates through precomputed lookup tables, which facilitate efficient processing. Although Heliosat-4 has been previously assessed in the region, as seen in studies like [11, 15], those evaluations did not include direct comparisons with other solar radiation modeling datasets. The data used in this work was obtained from the SoDa platform, with a temporal resolution of 15 minutes. The MSG satellite provides a spatial resolution of approximately 16 kilometers for this area.

#### 2.1.2 LSA-SAF MDSSFTD

The MSG Downwelling Surface Short-wave Radiation Fluxes – Total and Diffuse (MDSSFTD) product provides instantaneous estimates, updated every 15 minutes, of global and diffuse shortwave radiation reaching the Earth’s surface. The retrieval algorithm is divided into two distinct modules to handle clear-sky and cloudy-sky conditions separately [16, 17].

For clear-sky scenarios, the model employs the method outlined in [18], known as SIRAMix (Surface Incident Radiation using Aerosol Mixtures). This technique combines physical parameterizations with a precomputed lookup table (LUT) of aerosol optical properties, primarily containing values for direct and diffuse transmittance and corresponding aerosol albedos. The LUT is generated using radiative transfer simulations under varying conditions of aerosol concentration, water vapor, and solar zenith angle.

In the case of cloud cover, the retrieval of GHI relies on the total shortwave albedo of clouds at the top of the atmosphere (TOA), which is derived from reflectance measurements taken by MSG satellites. This cloudy-sky module uses the same aerosol inputs as the clear-sky module, while cloud contributions are modeled using a simplified radiative transfer approach similar to that in [19].

The final GHI values are calculated by combining the clear-sky transmittance—based on gases and aerosols—with the cloud transmittance estimated from the simplified model.

This dataset offers a spatial resolution of approximately 3 km and is updated every 15 minutes. Data can be accessed in NetCDF format at: <https://data.sasaf.lsasvcs.ipma.pt/PRODUCTS/MSG/MDSSFTD/>.

### 2.2 Machine Learning Models

Simple Linear Regression (SLR) is a fundamental statistical technique in machine learning and data analysis for modeling the relationship between a quantitative dependent variable and a single independent variable, which may be either quantitative or categorical after appropriate encoding. This method does not require the variables to be both continuous, nor does it assume that observations are equally spaced in time, allowing for application across diverse data types and contexts. The model estimates a straight-line relationship to predict the dependent variable based on the independent variable.

The core idea behind SLR is to find the best-fitting line—called the regression line—that minimizes the difference between the predicted and actual values. This is typically done using the method of least squares, which minimizes the sum of squared residuals (errors). The regression line is defined by two parameters: the slope, which represents the effect of the independent variable on the dependent variable,

and the intercept, which indicates the expected value of the dependent variable when the independent variable is zero.

The Multilayer Perceptron (MLP) [20] is a type of feedforward artificial neural network (ANN), meaning that information flows in a single direction—from the input layer to the output layer—through one or more intermediate layers known as hidden layers. In most cases, references to ANNs pertain specifically to MLPs. The basic architecture of an MLP consists of three types of layers: an input layer, one or more hidden layers, and an output layer. The nodes in the hidden and output layers are artificial neurons that employ nonlinear activation functions.

The MLP is trained using supervised learning techniques. The most commonly used training algorithm for MLPs is backpropagation, typically combined with an optimization algorithm. This learning procedure enables the MLP to minimize the discrepancy between its predicted outputs and the expected values, usually quantified by a loss function, over multiple training iterations or epochs.

Extreme Gradient Boosting (XGBoost), introduced by [21], is a powerful and widely-used machine learning technique known for its speed and accuracy. In regression tasks, XGBoost predicts continuous variables, such as temperature, by iteratively building decision trees and applying a boosting strategy to enhance prediction accuracy. Its main goal is to minimize prediction errors by combining several weak learners (decision trees) into a strong predictive model.

Understanding XGBoost requires familiarity with key concepts: decision trees, boosting, and the objective function. Decision trees split data based on feature-related decisions; boosting trains models in sequence, where each model aims to fix the mistakes of the previous ones. The objective function in XGBoost merges a loss function, which measures prediction error, with a regularization term that controls model complexity. The training process involves starting with a basic prediction, constructing trees to estimate residuals, updating predictions, adding new trees, and optimizing the model through regularization.

For both the MLP and XGBoost models, determining the optimal hyperparameter configuration was essential to achieve consistent performance across cross-validation folds. This was accomplished through an exhaustive grid search implemented with the GridSearchCV function from the Scikit-learn library in Python [22]. The specific hyperparameter ranges evaluated for the MLP and XGB models are described in 1. Selecting optimal hyperparameters is a critical step in machine learning, as they directly influence model complexity, generalization ability, and overall predictive performance [23].

Regarding input data preprocessing, normalization

Table 1: Cartesian hyperparameter space for supervised learning techniques.

Hyperparameter	Lower	Upper	By	Trans. function
<b>MLP</b>				
Hidden layers	1	3	1	-
Hidden nodes	1	4	1	$2^x$
Dropout fraction	0	0.3	0.1	-
Learning rate	-3	-1	1	$10^x$
<b>XGBoost</b>				
Booster	gbtree			
Estimators	1	50	10	-
Max. depth	2	5	1	$2^x$
Learning rate	-3	-1	1	$10^x$

or feature scaling is a common practice in machine learning to improve convergence and stability, particularly for models sensitive to feature magnitude. However, in this study, no normalization was applied, as it was not deemed necessary for the models and data used. For SLR, scaling of the independent variable is unnecessary because the model is inherently scale-invariant: multiplying the input by a constant factor results in an inverse proportional change in the slope coefficient, leaving predictions unchanged. Similarly, XGB, when implemented with decision trees as base learners, does not require feature scaling, since tree-based splits depend on the relative ordering of values rather than their absolute magnitudes [24, 21]. In the case of the MLP model, scaling can be important when input variables differ significantly in range or units. However, in the present study, a single input feature—satellite-derived GHI—was used, expressed in the same units and range as the target variable (ground-measured GHI). Therefore, additional normalization was not considered necessary in this case.

### 2.3 Ground Measurements

This study examines several measurement sites detailed in Table 2 and Figure 1, which include each site's code, geographic location, elevation, measurement timeframe, Köppen–Geiger climate classification [25]. These stations are located in northwestern Argentina and encompass diverse climatic and geographic settings, ranging from lowland subtropical regions to high-altitude Andean plateaus.

The Sa station, based on INENCO's experimental campus at the National University of Salta in the Lerma Valley, lies in a pre-Andean urban area where frequent cloud cover is influenced by nearby mountains. It features a Cwb climate (subtropical highland with dry winters and mild summers) and recorded data from 2013 to 2020 using an Eppley PSP pyranometer.

The Lq station in La Quiaca is located in a cold

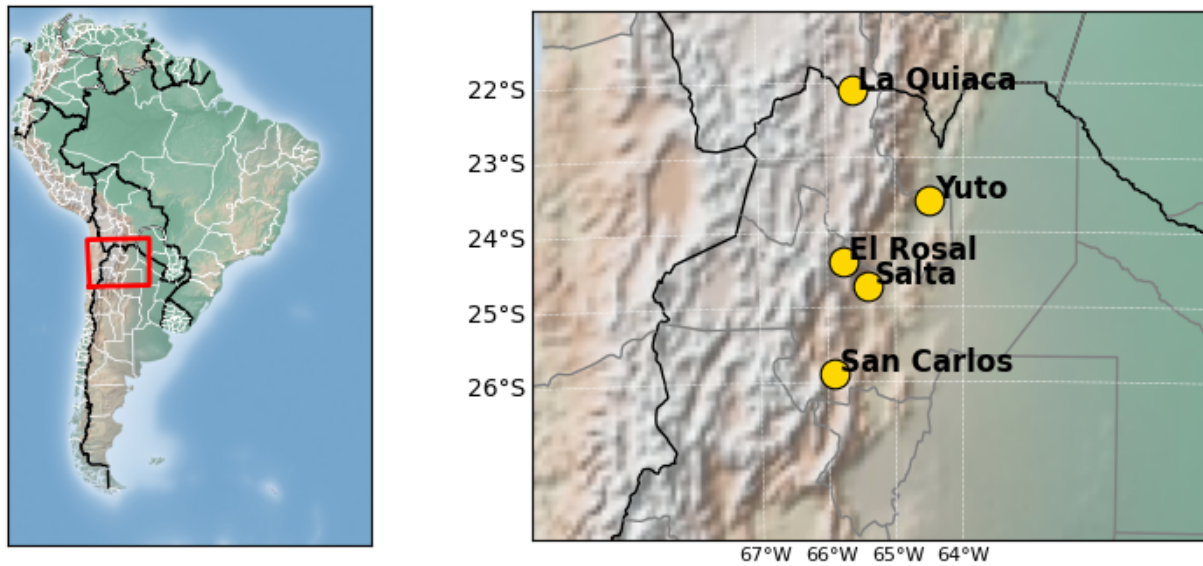


Figure 1: Map of South America showing a rectangle delineating the NOA region (Northwestern Argentina) on the left, and a detailed zoom of this region on the right with the cities Yuto, Salta, San Carlos, El Rosal, and La Quiaca marked with gold points and labeled. Both subplots display shaded relief and political boundaries.

semi-arid steppe (BSk) climate and is known for exceptionally high annual sunshine. Data from 2021 to 2023 were collected using a Kipp & Zonen CMP11 pyranometer.

The Yu station, set in a humid subtropical zone (Cwa), recorded measurements between 2017 and 2018 with a CMP11 sensor. The Sca station, also with a Cwb climate, gathered data from 2012 to 2013 using a CMP3 pyranometer. Lastly, the Ero station, located in a high-altitude BSk climate, used a CMP3 sensor to collect data from 2016 to 2018.

All locations utilized pyranometers certified to ISO 9060:2018 Class A or B standards for measuring global horizontal irradiance (GHI). Data were logged every minute, with each value averaging six instantaneous readings taken at 10-second intervals.

It is important to note that the dataset was divided into training, validation, and testing subsets following the same procedure described in [3]. Specifically, a one-year period of data was selected for model development, which was further split into 80% for training and 20% for validation. The remaining data was reserved exclusively for testing the models, ensuring that evaluation was performed on data not used during training or hyperparameter tuning.

The one-minute global horizontal irradiance (GHI) data were processed using a simplified quality control method based on the approach described in [26]. Since this study relies solely on GHI measurements and does not include diffuse irradiance, a reduced version of the original procedure was applied. Table 3 summarizes the filters used, where  $E$  is the solar constant,  $S$  the Earth–Sun distance correction factor,

$\theta_z$  the solar zenith angle, and  $kt$  the clearness index—defined as the ratio of GHI to the theoretical top-of-atmosphere irradiance on a horizontal plane.

The applied filters are as follows:

- F1: Rejects values exceeding a physically reasonable limit based on solar position.
- F2: Discards data points using a zenith-angle-dependent empirical upper threshold.
- F3: Filters out clearness index values greater than 1.4 when the sun is less than  $10^\circ$  above the horizon.

The percentage of daytime measurements retained varied among stations. Specifically, 73% of the data from Yu met the established criteria, compared to 82% in Sa, 72% in Sca, approximately 84% in Ero, and 69% in Lq.

After applying these quality control steps, one year of filtered hourly-averaged data was selected for model calibration, while the remaining observations were used for validation.

The ground-based measurement series used in this study for the Yu, Sa, Sca, and Ero stations are managed by the Grupo de Evaluación y Estudio del Recurso Solar (GEERS) of INENCO. These data can be requested by contacting [germansalazar@conicet.ar](mailto:germansalazar@conicet.ar). The measurements corresponding to the Lq station are managed and distributed by the Servicio Meteorológico Nacional Argentino (SMN) and can be obtained by contacting [cim@smn.gob.ar](mailto:cim@smn.gob.ar).

As a result of this procedure, all sites except Sca retained at least two years of valid data, from which the training, validation, and testing subsets were constructed. For these stations, the calibration and validation sets together cover a full annual cycle, and the test

Table 2: Stations considered in this work, with location, altitude, and climate classification.

ID	City	Latitude	Longitude	Altitude (m)	Climate
Yu	Yuto	-23.58	-64.50	401	Cwa
Sa	Salta	-24.72	-65.40	1233	Cwb
Sca	San Carlos	-25.89	-65.92	1624	Cwb
Ero	El Rosal	-24.39	-65.76	3355	BSk
Lq	La Quiaca	-22.10	-65.60	3468	BSk

Table 3: Quality control filters applied to the measurements.

Filter	Description
F1	$GHI < 1.5 E S (\cos(\theta_z))^{1.2} + 100 \text{ W/m}^2$
F2	$GHI > (6.5331 - 0.065502 \theta_z + 1.8312E-4 \theta_z^2) / (1 + 0.01113 \theta_z)$
F3	$kt < 1.4 \ \& \ (90 - \theta_z) < 10^\circ$

set also includes data from different seasons, ensuring a fair representation of seasonal variability. In contrast, the Sca station has a more limited measurement record: only 12 months of filtered data were available for training and validation, plus an additional 2 months for testing. Although this coverage still includes parts of the seasonal cycle, it does not encompass a complete second annual cycle in the test set, which is acknowledged as a limitation of this study.

### 3 Results

The most common performance indicators in the field of solar resource assessment have been covered by [27]; these include the Mean Bias Error (MBE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE). The three metrics are defined as follows,

$$\text{MBE} = \frac{\sum_{i=1}^n (y_i - x_i)}{n}, \quad (1)$$

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n}, \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2}, \quad (3)$$

where  $x$  and  $y$  are the measured and estimated values, respectively, and  $n$  is the sample size. The MBE measures the systematic bias that a model can introduce in a long-term evaluation, while the MAE and RMSE measure the dispersion of the error using absolute and quadratic norms, respectively. Because of its greater sensitivity to outliers, the RMSE is often used in this area. Both dispersion metrics are reported here for completeness. The three indicators are presented in relative terms as a percentage of the mean of the measured values, referred to here as MBE (%), MAE (%), RMSE (%).

The performance evaluation of the SLR, XGB, and MLP models, using input features from the CAMS and LSASAF satellite products, was conducted across five sites with diverse climatic and geographic characteristics. The results are expressed in terms of mean bias error (MBE), mean absolute error (MAE), and root mean square error (RMSE), all normalized with respect to the average GHI at each station (see Table 4).

Overall, normalized errors remain within a moderate range across all combinations of models and data sources, with relatively small differences between complex and linear approaches. The SLR model exhibited competitive performance on both datasets, often yielding error metrics that are close to those of the more sophisticated models.

For the CAMS dataset, while MLP achieved slightly lower RMSE values at some sites, it also introduced more pronounced biases in MBE, suggesting a trade-off between variance reduction and increased systematic error. XGB showed similar behavior to SLR, with only marginal differences.

Results using LSASAF indicated a slight overall improvement compared to CAMS, particularly at high-altitude stations (such as ERO and LQ), where both XGB and MLP achieved lower MAE and RMSE values. This improvement suggests that LSASAF's higher temporal resolution or enhanced cloud representation may offer more informative features for local site adaptation.

Nevertheless, no substantial gains were observed by increasing model complexity. These findings indicate that, given the current quality of input data, simple models like SLR are capable of capturing most of the relationship between satellite-based inputs and surface GHI measurements, while also providing greater robustness to data noise and training limitations.

The residual analysis shown in Figure 2 for the five study locations enables a comparative evaluation of the MLP and XGB as well as the two satellite-based input data sources: CAMS and LSA-SAF.

A systematic bias is evident across all locations, particularly when using CAMS data. Residuals exhibit a clear negative trend with respect to measured GHI, indicating that both models tend to overestimate irradiance under low GHI conditions and underestimate it at higher GHI levels. This behavior points to a

Table 4: Performance metrics (MBE, MAE, RMSE) for each model and satellite dataset across the five sites. Values are normalized and expressed as percentages relative to the mean GHI at each site: 429.2 W/m<sup>2</sup> (Yu), 432.9 W/m<sup>2</sup> (Sa), 554.4 W/m<sup>2</sup> (Sca), 628.4 W/m<sup>2</sup> (Ero), and 648.7 W/m<sup>2</sup> (Lq).

Model	Yu			Sa			Sca			Ero			Lq		
	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE
CAMS	≈ 0	15.9	25.1	3.6	21.1	29.6	3.6	24.2	31.8	-20.2	25.0	37.2	-6	14.2	21.6
SLR-CAMS	-0.6	15.4	24.2	-0.9	18.8	27.3	0.1	19.5	26.9	8.4	25.3	30.0	2.1	13.5	19.4
XGB-CAMS	-0.9	16.1	24.6	-0.9	18.9	27.4	0.2	19.9	27.2	8.1	24.0	29.4	2.3	13.6	19.5
MLP-CAMS	-3.0	15.7	24.5	-4.0	19.3	28.1	-5.5	21.6	29.3	-2.0	26.8	33.7	-2.1	14.3	20.8
LSASAF	7.3	15.4	24.7	16.1	24.3	34.2	16.4	26.1	34.4	-5.1	14.3	23.8	3.4	10.3	17.8
SLR-LSASAF	-5.4	17.3	24.1	-1.4	21.6	29.8	5.2	19.2	28.5	5.5	17.3	23.7	-0.6	10.6	17.4
XGB-LSASAF	-5.7	17.1	24.7	-1.4	21.5	29.9	3.9	20.6	29.3	5.6	17.0	23.5	-0.6	10.7	17.4
MLP-LSASAF	-4.1	17.6	24.1	-0.2	22.2	30.0	4.2	19.4	28.7	1.7	16.5	23.8	-1.1	10.7	17.4

limitation in the CAMS dataset's ability to accurately capture local irradiance variability. Conversely, the use of LSA-SAF data significantly reduces this bias, suggesting that this source provides a more accurate representation of local atmospheric conditions.

In addition, the discretization effect inherent to tree-based models is observed in the residuals generated by the XGB model. This manifests as horizontal banding patterns in the scatter plots, particularly when CAMS data is used. The effect is more pronounced at sites with higher irradiance variability, such as SA and SCA. This phenomenon is intrinsic to the decision-tree architecture of XGB, which yields predictions in discrete steps. In contrast, the MLP model produces continuous and smooth residual distributions, owing to its differentiable functional structure, and does not exhibit such discretization artifacts.

Overall, the LSA-SAF data source improves model performance by reducing systematic errors and better capturing site-specific atmospheric conditions. Additionally, the MLP model demonstrates increased robustness in producing smoother predictions, particularly in complex or highly variable environments where the limitations of tree-based discretization are more evident.

## 4 Discussion

The results from the site adaptation evaluations using two satellite-based datasets—CAMS and LSA-SAF—across different modeling approaches (SLR, XGB, and MLP) reveal several important insights regarding model complexity, data quality, and performance limitations.

Across all five sites (Yu, Sa, Sca, Ero, Lq), it is evident that increasing model complexity does not systematically lead to substantial improvements in error metrics. This is consistent in both datasets. For example, with CAMS inputs, while the MLP occasionally achieves slightly lower RMSE values (e.g., in Yu), its MBE is often worse (e.g., -3.0 in Yu, -5.5 in Sca), suggesting a trade-off between reducing variance and

increasing systematic bias. Similarly, in the LSA-SAF case, although XGB and MLP reduce MAE and RMSE in some stations (e.g., Ero and Lq), their performance across the board remains only marginally better than SLR, and occasionally worse.

This outcome suggests several possible causes:

**Ceiling Effect Due to Data Quality:** Both CAMS and LSA-SAF provide irradiance estimates derived from satellite observations and reanalysis data, which carry inherent uncertainties. The inputs may already contain biases or miss key high-resolution phenomena (e.g., microclimate effects, local shading, cloud variability). Therefore, even complex models like MLP or XGB have limited room for learning meaningful corrections beyond what simpler models can already capture.

**Redundancy in Model Capacity:** When the predictive relationship between the satellite product and ground truth GHI is primarily linear or moderately nonlinear, simpler models like SLR are often sufficient. In such cases, adding complexity (e.g., more parameters or nonlinear transformations) may not translate to significantly better predictions, especially if the noise level in the data is high relative to the signal.

**Model Overfitting and Generalization:** The MLP model, being the most flexible and prone to overfitting, performs inconsistently. It performs well in certain metrics (e.g., RMSE in YU with LSA-SAF) but exhibits considerable bias in others (e.g., SCA with CAMS, where MBE is -5.5). These discrepancies may reflect poor generalization due to limited or noisy training samples.

**Dataset Characteristics and Station Dependency:** Interestingly, LSA-SAF-based models show slightly lower MAE and RMSE in most stations (especially Ero and Lq), with MLP and XGB performing notably better than their CAMS counterparts. This suggests that the LSA-SAF dataset, perhaps due to its higher temporal resolution or improved cloud property retrieval, provides slightly more informative features for site adaptation.

**Robustness of Simple Models:** The SLR model con-

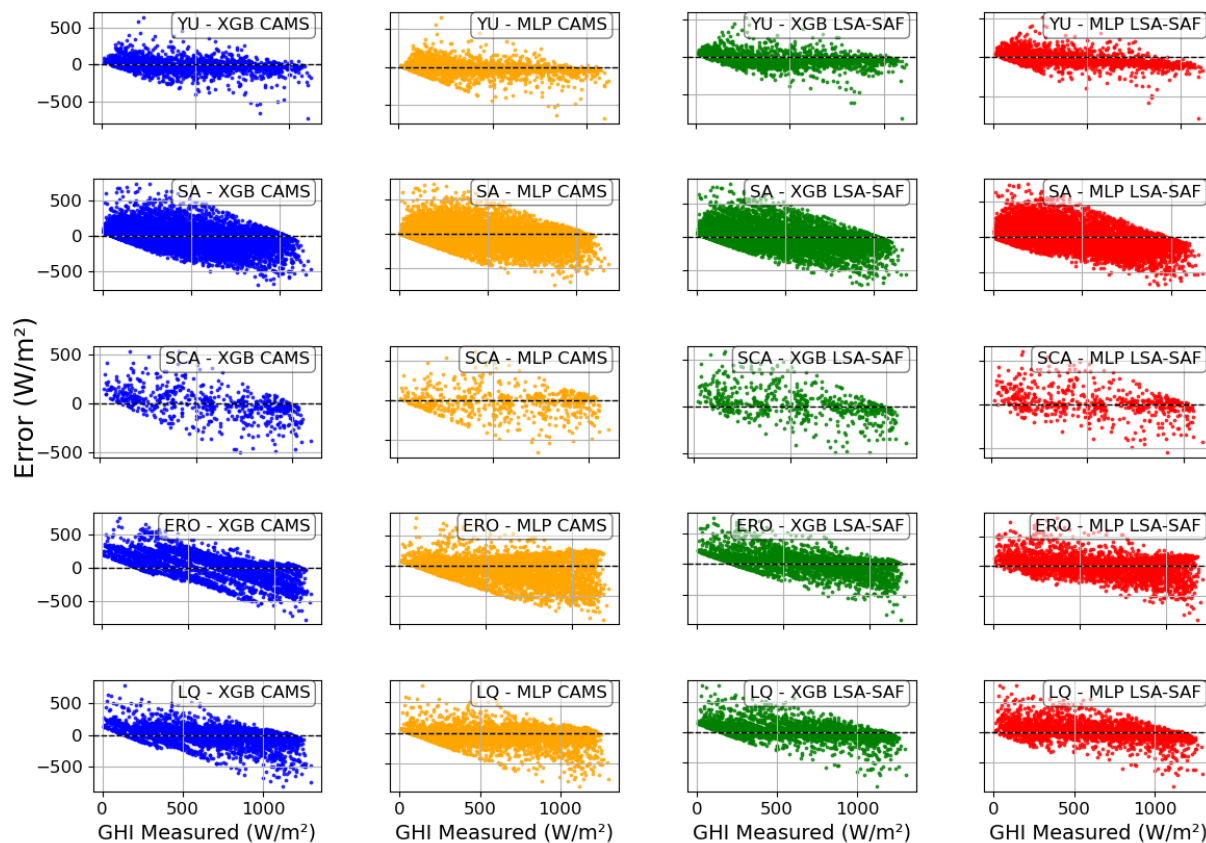


Figure 2: Residual plots of GHI predictions for five locations (YU, SA, SCA, ERO, and LQ) using two machine learning models (XGB and MLP) and two satellite-based input datasets (CAMS and LSA-SAF). Each subplot shows the prediction error (in  $W/m^2$ ) as a function of measured GHI. Systematic biases and model-specific behaviors—such as over/underestimation trends and discretization effects—are visually identifiable across different configurations.

sistently performs competitively in both datasets, often yielding RMSEs within 1 unit of more complex models. This underscores the robustness and interpretability of linear approaches, especially when data limitations dominate the modeling landscape.

In conclusion, these findings indicate that under current data conditions, the potential performance improvements from employing complex machine learning models are modest. Enhancing the quality and resolution of the input data—e.g., incorporating high-resolution meteorological or topographic information—may yield greater benefits than further increasing model complexity. For practical applications where computational efficiency is a concern, simpler models like SLR may still offer the most balanced solution.

One limitation of the present work concerns the Sca site, for which only 14 months of filtered data were available. While the calibration period covers a complete annual cycle, the test set comprises only two additional months, reducing the representation of interannual variability at this location. This constraint may affect the generalization assessment for Sca more than for the other sites, where multi-year records were available. Nevertheless, the consistency of the results

across stations with longer datasets supports the robustness of the main conclusions.

#### 4.1 Potential Improvements and Future Directions

Given the limited performance gains observed with more complex models, future work should explore enhancements not only on the algorithmic side but also in the quality, diversity, and structure of the input data. Several strategies may be proposed to improve irradiance estimation accuracy in site-adapted models:

**Incorporation of Additional Meteorological and Environmental Variables:** The current models rely heavily on satellite-derived irradiance inputs (e.g., CAMS or LSA-SAF), which may not fully capture local conditions. Incorporating complementary variables—such as temperature, relative humidity, wind speed, aerosol optical depth, or cloud cover metrics—could help the models better understand physical mechanisms affecting GHI. Including ground-based measurements, if available, would be even more beneficial.

**Temporal Decomposition: Trend and Seasonality Analysis:** Irradiance time series often exhibit strong diurnal and seasonal cycles. Incorporating explicit

temporal features—such as time of day, day of year, or periodicity terms—can help models capture these patterns more effectively. Decomposing the signal into trend, seasonal, and residual components, and modeling them separately (e.g., through additive models or hybrid statistical-ML frameworks), could lead to more accurate forecasts.

**Cluster-Based Regression Models:** Site-specific characteristics such as altitude, terrain, and microclimate variability may produce heterogeneous relationships between inputs and outputs. Applying clustering techniques (e.g., k-means, DBSCAN) to group observations with similar patterns, and then training specialized regression models per cluster, could allow better adaptation to localized dynamics. Alternatively, clustering based on weather regimes or cloud types may also be explored.

**Model Ensemble and Hybrid Approaches:** While individual models may have limitations, combining multiple models (e.g., ensemble of SLR, XGB, and MLP) may yield improved robustness and accuracy. Hybrid approaches that integrate physical models (e.g., clear-sky models) with data-driven corrections could also provide better interpretability and performance.

**Time-Series Models and Sequential Learning:** The current models treat irradiance estimation as a static regression problem. However, time dependencies could be leveraged through autoregressive methods, recurrent neural networks (e.g., LSTM), or Transformer-based architectures. These models could capture temporal persistence, short-term fluctuations, and lagged effects that are missed by purely static approaches.

**Error Correction Techniques:** Post-processing methods, such as bias correction or quantile mapping, can help correct systematic errors in model outputs. Applying these methods on residuals or forecast errors may improve calibration and reduce bias without retraining the core models.

**Use of High-Resolution or Localized Data Sources:** The spatial resolution of datasets like CAMS and LSA-SAF may not be sufficient to resolve terrain-induced variability in solar radiation. Incorporating higher-resolution satellite data (e.g., from Sentinel-2, Himawari, or GOES), or downscaling techniques, could help improve local estimates. Topographic corrections using digital elevation models (DEM) may also be considered.

**Model Explainability and Error Analysis:** A deeper analysis of residuals by season, time of day, or cloud type may reveal when and where each model fails, guiding future enhancements. Explainable AI (XAI) tools like SHAP or LIME could be applied to understand variable importance and model decisions.

These directions point toward a more holistic and nuanced approach to irradiance modeling, moving beyond one-size-fits-all regression and toward context-aware, physically-informed, and temporally-sensitive strategies.

## 5 Conclusions

The findings of this study indicate that, in the context of site adaptation for global horizontal irradiance (GHI) using satellite-derived products such as CAMS and LSA-SAF, the evaluated machine learning models—Simple Linear Regression (SLR), Extreme Gradient Boosting (XGB), and Multilayer Perceptron (MLP)—exhibited comparable performance in terms of standard error metrics, with only marginal differences among them. Despite its simplicity, SLR demonstrated a remarkable ability to capture the relationship between satellite-derived data and ground measurements, achieving relative error levels similar to those obtained by the more complex XGB and MLP models.

These results suggest that the quality of the input data imposes an upper limit on the achievable accuracy improvements offered by complex models. The satellite products employed contain inherent uncertainties and lack the resolution required to account for localized, high-frequency phenomena such as microclimatic effects or terrain-induced variability. Consequently, increasing model complexity does not substantially enhance predictive performance under these data conditions. In this regard, the predictive capacity of complex models becomes redundant when the underlying relationship between variables is predominantly linear or only mildly nonlinear, which explains the strong performance of SLR.

Furthermore, the analysis revealed that more flexible models, such as MLP, are prone to overfitting, leading to inconsistent behavior across different performance metrics. This underscores the importance of balancing model complexity with data quality to avoid compromising the model's generalization ability.

In summary, given the current data quality and site characteristics, simple models such as SLR represent a robust, interpretable, and computationally efficient solution for site-adapted GHI estimation. Significant improvements in estimation accuracy are more likely to be achieved by incorporating additional meteorological variables, applying temporal decomposition techniques, or developing hybrid approaches that combine physical models with statistical corrections, rather than by merely increasing the complexity of machine learning algorithms.

### Competing interests

The authors have declared that no competing interests exist.

### Authors' contribution

The authors confirm contribution to the paper as follows: RL: Conceptualization, Methodology, Software, Writing-Original draft preparation; GS: Writing-Reviewing and Editing.

## References

- [1] R. Alonso-Suárez, “Estimación del recurso solar en uruguay mediante imágenes satelitales,” Ph.D. dissertation, Universidad de la República - Uruguay, 07 2017.
- [2] H. Grossi Gallegos and R. Righini, *Atlas de Energía Solar de la República Argentina*, 05 2007.
- [3] G. Salazar, R. D. Ledesma, C. López Ruiz, and O. de Castro Vilela, “Análisis de desempeño de diferentes técnicas de aprendizaje automático en una adaptación al sitio de irradiancia solar global para salta (argentina),” *Avances en Energías Renovables y Medio Ambiente - AVERMA*, vol. 28, p. 308–320, abr. 2025. [Online]. Available: <https://portalderevistas.unsa.edu.ar/index.php/averma/article/view/4892>
- [4] R. Laspiur, G. A. Salazar, J. Zerpa, and M. Watkins, “Trazado de mapas medios anuales de energía solar global, directa, difusa y tilt, usando la base de datos de swera. caso de estudio: provincias de salta y jujuy,” *Avances en Energías Renovables y Medio Ambiente - AVERMA*, vol. 17, p. 47–52, nov. 2021. [Online]. Available: <https://portalderevistas.unsa.edu.ar/index.php/averma/article/view/2066>
- [5] N. Sarmiento, S. Belmonte, P. Dellicompagni, J. Franco, K. Escalante, and J. Sarmiento, “A solar irradiation gis as decision support tool for the province of salta, argentina,” *Renewable Energy*, vol. 132, pp. 68–80, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960148118308784>
- [6] J. Polo, S. Wilbert, J. Ruiz-Arias, R. Meyer, C. Gueymard, M. Súrri, L. Martín, T. Mieslinger, P. Blanc, I. Grant, J. Boland, P. Ineichen, J. Remund, R. Escobar, A. Troccoli, M. Sengupta, K. Nielsen, D. Renne, N. Geuder, and T. Cebecauer, “Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets,” *Solar Energy*, vol. 132, pp. 25–37, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0038092X16001754>
- [7] J. Polo, C. Fernández-Peruchena, V. Salamalikis, L. Mazorra-Aguilar, M. Turpin, L. Martín-Pomares, A. Kazantzidis, P. Blanc, and J. Remund, “Benchmarking on improvement and site-adaptation techniques for modeled solar radiation datasets,” *Solar Energy*, vol. 201, pp. 469–479, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0038092X20302784>
- [8] H. Verbois, Y.-M. Saint-Drenan, Q. Libois, Y. Michel, M. Cassas, L. Dubus, and P. Blanc, “Improvement of satellite-derived surface solar irradiance estimations using spatio-temporal extrapolation with statistical learning,” *Solar Energy*, vol. 258, pp. 175–193, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0038092X23002670>
- [9] V. Salamalikis, P. Tzoumanikas, A. A. Argiriou, and A. Kazantzidis, “Site adaptation of global horizontal irradiance from the copernicus atmospheric monitoring service for radiation using supervised machine learning techniques,” *Renewable Energy*, vol. 195, pp. 92–106, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960148122008758>
- [10] S. Zainali, D. Yang, T. Landelius, and P. E. Campana, “Site adaptation with machine learning for a northern europe gridded global solar irradiance product,” *Energy and AI*, vol. 15, p. 100331, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666546823001039>
- [11] G. A. Salazar, R. Alonso-Suárez, A. Laguarda Cirigliano, and R. D. Ledesma, “Evaluación del proceso de adaptación al sitio aplicado a la irradiancia solar global medida en la ciudad de salta, argentina,” *Avances en Energías Renovables y Medio Ambiente - AVERMA*, vol. 25, p. 353–362, may 2022. [Online]. Available: <https://portalderevistas.unsa.edu.ar/index.php/averma/article/view/2431>
- [12] Z. Qu, A. Oumbe, P. Blanc, B. Espinar, G. Gesell, B. Gschwind, L. Klüser, M. Lefèvre, L. Saboret, M. Schroedter-Homscheidt, and L. Wald, “Fast radiative transfer parameterisation for assessing the surface solar irradiance: The heliosat24 method,” *Meteorologische Zeitschrift*, vol. 26, no. 1, pp. 33–57, 02 2017. [Online]. Available: <http://dx.doi.org/10.1127/metz/2016/0781>
- [13] B. Mayer and A. Kylling, “Technical note: The libradtran software package for radiative transfer calculations - description and examples of use,” *Atmospheric Chemistry and Physics*, vol. 5, no. 7, pp. 1855–1877, 2005. [Online]. Available: <https://acp.copernicus.org/articles/5/1855/2005/>
- [14] K. T. Kriebel, G. Gesell, M. Ka’stner, and H. M. and, “The cloud analysis tool apollo: Improvements and validations,” *International Journal of Remote Sensing*, vol. 24, no. 12, pp. 2389–2408, 2003. [Online]. Available: <https://doi.org/10.1080/01431160210163065>
- [15] R. Ledesma, G. Salazar, and O. Vilela, “Avances en la estimación de irradiancia solar en las provincias de salta y jujuy mediante imágenes satelitales goes-16,” *Avances en Energías Renovables y Medio Ambiente - AVERMA*, vol. 27, p. 413–427, sep. 2024. [Online]. Available: <https://portalderevistas.unsa.edu.ar/index.php/averma/article/view/4643>
- [16] M. Derrien and H. L. Gléau, “MSG/SEVIRI cloud mask and type from SAFNWC,” *International Journal of Remote Sensing*, vol. 26, no. 21, pp. 4707–4732, 2005, available: <https://doi.org/10.1080/01431160500166128>. [Online]. Available: <https://doi.org/10.1080/01431160500166128>
- [17] N. SAF, “User manual for the cloud product processors of the nwc/geo: Science part,” *Météo-France / Centre d’études en Météorologie Satellitaire*, Tech. Rep., 2022.
- [18] X. Ceamanos, D. Carrer, and J.-L. Roujean, “Improved retrieval of direct and diffuse downwelling surface shortwave flux in cloudless atmosphere using dynamic estimates of aerosol content and type: application to the LSA-SAF project,” *Atmospheric Chemistry and Physics*, vol. 14, no. 15, pp. 8209–8232, 2014, available: <https://doi.org/10.5194/acp-14-8209-2014>. [Online]. Available: <https://acp.copernicus.org/articles/14/8209/2014/>

- [19] B. Geiger, C. Meurey, D. Lajas, L. Franchistéguy, D. Carrer, and J.-L. Roujean, “Near real-time provision of downwelling shortwave radiation estimates derived from satellite observations,” *Meteorological Applications*, vol. 15, no. 3, pp. 411–420, 2008, available: <https://doi.org/10.1002/met.84>. [Online]. Available: <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.84>
- [20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. [Online]. Available: <https://doi.org/10.1038/323533a0>
- [21] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: <https://doi.org/10.1145/2939672.2939785>
- [22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, and G. Louppe, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, 01 2012.
- [23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, book in preparation for MIT Press. [Online]. Available: <http://www.deeplearningbook.org>
- [24] E. Soria Olivas, P. Rodríguez Belenguer, Q. García Vidal, F. Vaquer Estalrich, J. V. Camisón, and J. Vila Tomás, *Inteligencia artificial: Casos prácticos con aprendizaje profundo*, 1st ed. Bogotá, Colombia: Ediciones de la U, 2022.
- [25] M. C. Peel, B. L. Finlayson, and T. A. McMahon, “Updated world map of the köppen-geiger climate classification,” *Hydrology and Earth System Sciences*, vol. 11, no. 5, pp. 1633–1644, 2007. [Online]. Available: <https://hess.copernicus.org/articles/11/1633/2007/>
- [26] F. M. Nollas, G. A. Salazar, and C. A. Gueymard, “Quality control procedure for 1-minute pyranometric measurements of global and shadowband-based diffuse solar irradiance,” *Renewable Energy*, vol. 202, pp. 40–55, 2023, available: <https://doi.org/10.1016/j.renene.2022.11.056>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960148122016962>
- [27] J. Zhang, A. Florita, B.-M. Hodge, S. Lu, H. F. Hamann, V. Banunarayanan, and A. M. Brockway, “A suite of metrics for assessing the performance of solar power forecasting,” *Solar Energy*, vol. 111, pp. 157–175, 2015, available: <https://doi.org/10.1016/j.solener.2014.10.016>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0038092X14005027>

**Citation:** R. Ledesma and G. Salazar. *Machine Learning for Site Adaptation of Satellite-Derived Solar Irradiance in Northwestern Argentina*. Journal of Computer Science & Technology, vol. 25, no. 2, pp. 118–127, 2025.  
**DOI:** 10.24215/16666038.25.e10.  
**Received:** June 18, 2025 **Accepted:** August 30, 2025.  
**Copyright:** This article is distributed under the terms of the Creative Commons License CC-BY-NC-SA.